6.2.4 Types of stimulus-comparison methods

Three types of stimulus-comparison methods have been used in television assessments.

6.2.4.1 Adjectival categorical judgement methods

In adjectival categorical judgement methods, observers assign the relation between members of a pair to one of a set of categories that, typically, are defined in semantic terms. These categories may report the existence of perceptible differences (e.g. SAME, DIFFERENT), the existence and direction of perceptible differences (e.g. LESS, SAME, MORE), or judgements of extent and direction. The ITU-R comparison scale is shown in Table 4 below.

TABLE 4

Comparison scale

–3	Much worse
–2	Worse
–1	Slightly worse
0	The same
+1	Slightly better
+2	Better
+3	Much better

This method yields a distribution of judgements across scale categories for each condition pair. The way that responses are analysed depends on the judgement made (e.g. difference) and the information required (e.g. just-noticeable differences, ranks of conditions, "distances" among conditions, etc.).

6.2.4.2 Non-categorical judgement methods

In non-categorical judgements, observers assign a value to the relation between the members of an assessment pair. There are two forms of this method:

- In continuous scaling, the assessor assigns each relation to a point on a line drawn between two labels (e.g. SAME-DIFFERENT or the ends of a categorical scale as in Table 4). Scales may include additional reference labels at intermediate points. The distance from one end of the line is taken as the value for each condition pair.
- In the second form, the assessor assigns each relation a number that reflects its judged level on a specified dimension (e.g. difference in quality). The range of numbers used may be constrained or not. The number assigned may describe the relation in "absolute" terms or in terms of that in a "standard" pair.

Both forms result in a distribution of values for each pair of conditions. The method of analysis depends on the nature of the judgement and the information required.

6.2.4.3 Performance methods

In some cases, performance measures can be derived from stimulus-comparison procedures. In the forced-choice method, the pair is prepared such that one member contains a particular level of an attribute (e.g. impairment) while the other contains either a different level or none of the attribute. The observer is asked to decide either which member contains the greater/lesser level of the attribute or which contains any of the attribute; accuracy and speed of performance are taken as indices of the relation between the members of the pair.

6.3 Remarks

Other techniques, like multi-dimensional scaling methods and multivariate methods, are described in Report ITU-R BT.1082-1, and are still under study.

All of the methods described so far have strengths and limitations and it is not yet possible to definitively recommend one over the others. Thus, it remains at the discretion of the researcher to select the methods most appropriate to the circumstances at hand.

The limitations of the various methods suggest that it may be unwise to place too much weight on a single method. Thus, it may be appropriate to consider more "complete" approaches such as either the use of several methods or the use of the multi-dimensional approach.

APPENDIX 1

TO ANNEX 1

Picture-content failure characteristics

1 Introduction

Following its implementation, a system will be subjected to a potentially broad range of programme material, some of which it may be unable to accommodate without loss in quality. In considering the suitability of the system, it is necessary to know both the proportion of programme material that will prove critical for the system and the loss in quality to be expected in such cases. In effect, what is required is a picture-content failure characteristic for the system under consideration.

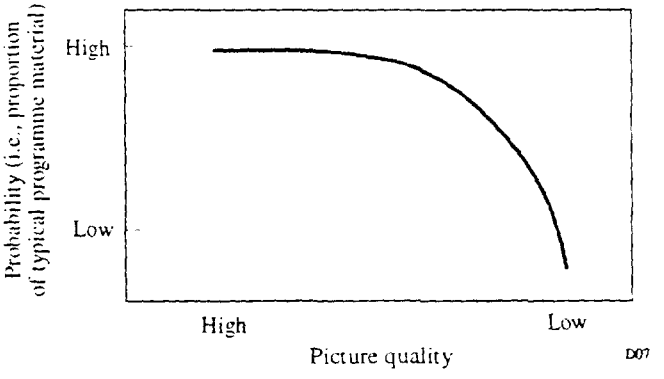
Such a failure characteristic is particularly important for systems whose performance may not degrade uniformly as material becomes increasingly critical. For example, certain digital and adaptive systems may maintain high quality over a large range of programme material, but degrade outside this range.

2 Deriving the failure characteristic

Conceptually, a picture-content characteristic establishes the proportion of the material likely to be encountered in the long run for which the system will achieve particular levels of quality. This is illustrated in Fig. 7.

FIGURE 7

Graphical representation of possible picture-content failure characteristic



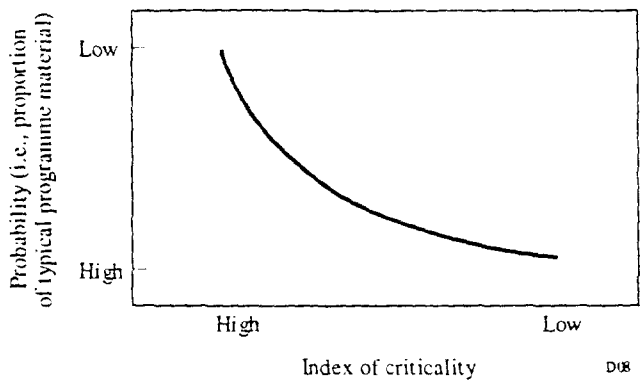
A picture-content failure characteristic may be derived in four steps:

Step 1 involves the determination of an algorithmic measure of "criticality" which should be capable of ranking a number of image sequences, which have been subjected to distortion from the system or class of systems concerned, in such a way that the rank order corresponds to that which would be obtained had human observers performed the task. This criticality measure may involve aspects of visual modelling.

Step 2 involves the derivation, by applying the criticality measure to a large number of samples taken from typical television programmes, of a distribution that estimates the probability of occurrence of material which provides different levels of criticality for the system, or class of systems, under consideration. An example of such a distribution is illustrated in Fig. 8.

FIGURE 8

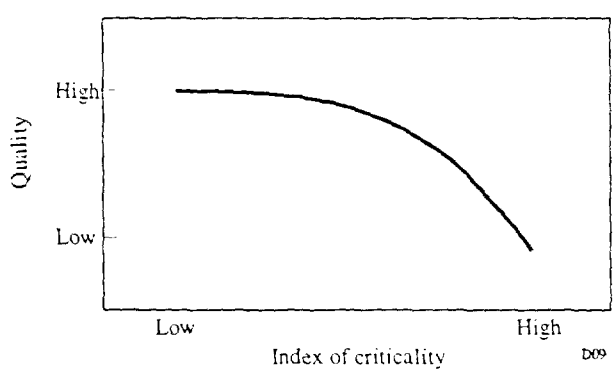
Probability of occurrence of material of differing levels of criticality



Step 3 involves the derivation, by empirical means, of the ability of the system to maintain quality as the level of criticality of programme material is increased. In practice, this requires subjective assessment of the quality achieved by the system with material selected to sample the range of criticality identified in Step 2. This results in a function relating the quality achieved by the system to the level of criticality in programme material. An example of such a function is given in Fig. 9.

FIGURE 9

A possible function relating quality to the criticality of programme material



Step 4 involves the combination of information from Steps 2 and 3 in order to derive a picture-content failure characteristic of the form given in Fig. 7.

3 Use of the failure characteristic

In providing an overall picture of the performance likely to be achieved over the range of possible programme material, the failure characteristic is an important tool for considering the suitability of systems. The failure characteristic can be used in three ways:

- to optimize parameters (e.g. source resolution, bit rate, bandwidth) of a system at the design stage to match it more closely to the requirements of a service;
- to consider the suitability of a single system (i.e. to anticipate the incidence and severity of failure during operation);
- to assess the relative suitabilities of alternative systems (i.e. to compare failure characteristics and determine which system would be more suitable for use). It should be noted that, while alternative systems of a similar type may use the same index of criticality, it is possible that systems of a dissimilar type may have different indices of criticality. However, as the failure characteristic expresses only the probability that different levels of quality will be seen in practice, characteristics can be compared directly even when derived from different, system-specific indices of criticality.

While the method described in this Recommendation provides a means of measuring the picture-content failure characteristic of a system, it may not fully predict the acceptability of the system to the viewer of a television service. To obtain this information it may be necessary for a number of viewers to watch programmes encoded with the system of interest, and to examine their comments.

APPENDIX 2

TO ANNEX 1

Method of determining a composite failure characteristic for programme content and transmission conditions

1 Introduction

A composite failure characteristic relates perceived image quality to probability of occurrence in practice in a way that explicitly considers both programme content and transmission conditions.

In principle, such a characteristic could be derived from a subjective study that involves sufficient numbers of observations, times of test, and reception points to yield a sample that represents the population of possible programme content and transmission conditions. In practice, however, an experiment of this sort may be impracticable.

The present Appendix describes an alternative, more readily realized procedure for determining composite failure characteristics. This method consists of three stages:

- programme-content analysis;
- transmission-channel analysis; and
- derivation of composite failure characteristics.

2 Programme-content analysis

This stage involves two operations. First, an appropriate measure of programme content is derived and, second, the probabilities with which values of this measure occur in practice are estimated.

A programme-content measure is a statistic that captures aspects of programme content that stress the ability of the system(s) under consideration to provide perceptually faithful reproductions of programme material. Clearly, it would be advantageous if this measure were based on an appropriate perceptual model. However, in the absence of such a model, a measure that captures some aspect of the extent of spatial diversity within and across video frames/fields might suffice, provided this measure enjoys a roughly monotonic relation with perceived image quality. It may be necessary to use different measures for systems (or classes of systems) that use fundamentally different approaches to image representation.

Once an appropriate measure has been selected, it is necessary to estimate the probabilities with which the possible values of this statistic occur. This can be done in one of two ways:

- with the empirical procedure, a random sample of perhaps 200 10 s programme segments in a studio format suited in resolution, frame rate, and aspect ratio to the system(s) considered is analysed. Analysis of this sample yields relative frequencies of occurrence for values of the statistic which are taken as estimates of probability of occurrence in practice: or
- with the theoretical method, a theoretical model is used to estimate the probabilities. It should be noted that, although the empirical method is preferred, it may be necessary in specific cases to use the theoretical method (e.g. when there is not sufficient information about programme content, such as with the emergence of new production technologies).

The foregoing analyses will result in a probability distribution for values of the content statistic (see also Appendix 1). This will be combined with the results of the transmission-conditions analysis to prepare for the final stage of the process.

3 Transmission-channel analysis

This stage also involves two operations. First, a measure of transmission-channel performance is derived. And, second, the probabilities with which values of this measure occur in practice are estimated.

A transmission-channel measure is a statistic that captures aspects of channel performance that influence the ability of the system(s) under consideration to provide perceptually faithful reproductions of source material. Clearly, it would be advantageous if this measure were based on an appropriate perceptual model. However, in the absence of such a model, a measure that captures some aspect of the stress imposed by the channel might suffice, provided this measure enjoys a roughly monotonic relation with perceived image quality. It may be necessary to use different measures for systems (or classes of systems) that use fundamentally different approaches to channel coding.

Once an appropriate measure has been selected, it is necessary to estimate the probabilities with which the possible values of this statistic occur. This can be done in one of two ways:

- with the empirical procedure, channel performance is measured at perhaps 200 randomly selected times and reception points. Analysis of this sample yields relative frequencies of occurrence for values of the statistic which are taken as estimates of probability of occurrence in practice; or
- with the theoretical method, a theoretical model is used to estimate the probabilities. It should be noted that, although the empirical method is preferred, it may be necessary in specific cases to use the theoretical method (e.g. when there is not sufficient relevant information about channel performance, such as with the emergence of new transmission technologies).

The foregoing analyses will result in a probability distribution for values of the channel statistic. This will be combined with the results of the programme-content analysis to prepare for the final stage of the process.

4 Derivation of composite failure characteristics

This stage involves a subjective experiment in which programme content and transmission conditions are varied jointly according to probabilities established in the first two stages.

The basic method used is the double-stimulus continuous quality procedure and, in particular, the 10 s version recommended for motion sequences (see Annex 1, § 5). Here, the reference is a picture at studio quality in an appropriate format (e.g. one with resolution, a frame rate, and an aspect ratio appropriate to the system(s) considered). In contrast, the test presents the same picture as it would be received in the system(s) considered under selected channel conditions.

Test material and channel conditions are selected in accordance with probabilities established in the first two stages of the method. Segments of test material, each of which has been analysed to determine its predominant value according to the content statistic, comprise a selection pool. Material is then sampled from this pool such that it covers the range of possible values of the statistic, sparsely at less critical levels and more densely at more critical levels. Possible values of the channel statistic are selected in a similar way. Then, these two independent sources of influence are combined randomly to yield combined content and channel conditions of known probability.

The results of such studies, which relate perceived image quality to probability of occurrence in practice, are then used to consider the suitability of a system or to compare systems in terms of suitability.

ANNEX 2

Analysis and presentation of results

1 Introduction

In the course of a subjective experiment to assess the performance of a television system, a large amount of data is collected. These data, in the form of observers' score sheets, or their electronic equivalent, must be condensed by statistical techniques to yield results in graphical and/or numerical form which summarize the performance of the systems under test.

The following analysis is applicable to the results of the double stimulus impairment scale (DSI) method and the double stimulus continuous quality scale (DSCQ) method for the assessments of television picture quality which are found in Annex 1 (§ 4 and 5). In the first case, the impairment is rated on a five-point scale. In the second case, continuous rating scales are used and the results (differences of the ratings for the reference picture and the actual picture under test) are normalized to integer values between 0 and 100.

2 Common methods of analysis

The tests carried out according to the principles of the DSI or the DSCQ method lead to a certain distribution of integer values between 0 and 5 or between 0 and 100. This distribution includes the differences in judgement between observers and the effect of a variety of conditions associated with the experiment, for example the use of several pictures.

2.1 Calculation of mean scores

The first step of the analysis of the results is the calculation of the mean score for each test condition:

$$U = \frac{1}{N} \sum_{i=1}^{N} u_i \tag{1}$$

u_i: score of observer i

N: number of observers.

2.2 Calculation of confidence region

Even for objective measurements, the reliability of results is generally indicated by means of standard deviation. Knowing the strong standard deviation reported for individual subjective assessments, many observations are needed and the correct information about reliability is not the standard deviation but the confidence interval.

It is proposed to use the 5% confidence interval which is given by:

$$[u - \delta, u + \delta]$$

where:

$$\delta = 1.96\ S/\sqrt{N} \tag{2}$$

S: standard deviation

N: number of observers.

The standard deviation is provided by:

$$S = \sqrt{\sum_{i=1}^{N} (U - U_i)^2/(N - 1)} \tag{3}$$

With a probability of 95%, the absolute value of the difference between the experimental mean score and the "true" mean score (for a very high number of observers) is smaller than the 5% confidence interval, on condition that the distribution of the individual scores meets certain requirements.

2.3 Screening of the observers

First, it must then be ascertained whether this distribution of scores for each scene is normal or not using the β_2 test (by calculating the kurtosis coefficient of the function, i.e. the ratio of the fourth order moment to the square of the second order moment). If β_2 is between 2 and 4, the distribution may be taken to be normal. The scores U_i of each distribution i must then be compared with the associated mean value U plus the associated standard deviation times two (if normal) or times $\sqrt{20}$ (if non-normal), P_i, and to the associated mean value minus the same standard deviation times two or times $\sqrt{20}$, Q_i. Every time an observer's score is found above or below this range, this must be registered on a counter associated with each observer; two separate counters should be used for values above (P_i) and below (Q_i). Finally, the

following two ratios must be calculated: $P_i + Q_i$ over the total number of scores from each observer for the whole session, and $P_i - Q_i$ over $P_i + Q_i$ as an absolute value. If the former is greater than 5% and the latter less than 30%, observer i must be eliminated (see Note 1).

NOTE 1 – This procedure should not be applied more than once to the results of a given experiment. Moreover, use of the procedure should be restricted to cases in which there are relatively few observers (e.g., fewer than 20), all of whom are non-experts.

For each scene the above procedure can also be expressed mathematically as:

If $U_i > U + c\,S$ then $P_i = P_i + 1$

If $U_i > U - c\,S$ then $Q_i = Q_i + 1$

where $c = 2$ in the case of a normal distribution and $c = \sqrt{20}$ otherwise.

$$\text{If } \frac{P_i + Q_i}{\text{Total score for observer}} > 0.05 \text{ and } \left|\frac{P_i - Q_i}{P_i + Q_i}\right| < 0.3$$

then reject observer i.

This procedure is recommended for the EBU method; it has also been successfully applied to the double-stimulus continuous quality-scale method.

3 Processing to find a relationship between the mean score and the objective measure of a picture distortion

If subjective tests were carried out using the DSI method in order to investigate the relation between the objective measure of a distortion D and the mean score U, the following process can be useful, which consists of finding a simple continuous relationship between U and D.

3.1 Approximation by a symmetrical logistic function

The approximation of this experimental relationship by a logistic function is particularly interesting.

The processing of the data U can be made as follows:

The scale of values U is normalized by taking a continuous variable u such that:

$$u = (U - U_{min}) / (U_{max} - U_{min}) \tag{4}$$

when U is in the range U_{min} to U_{max}.

Graphical representation of the relationship between u and D shows that the curve tends to be a skew-symmetrical sigmoid shape provided that the natural limits to the values of D extend far enough from the region in which u varies rapidly.

The function $u = f(D)$ can now be approximated by a judiciously chosen logistic function, as given by the general relation:

$$u = \frac{1}{1 + e^{(D - D_M)G}} \tag{5}$$

where D_M and G are constants and G may be positive or negative.

The value u obtained from the optimum logistic function approximation is used to provide a deduced numerical value I according to the relation:

$$I = (1/u) - 1 \tag{6}$$

The values of D_M and G can be derived from the experimental data after the following transformation:

$$I = e^{(D - D_M)G} \tag{7}$$

This yields a linear relation by the use of a logarithmic scale for I:

$$\log_e I = (D - D_M) G$$

Interpolation by a straight line is simple and in some cases of an accuracy which is sufficient for the straight line to be considered as representing the impairment due to the effect measured by D.

The slope of the characteristic is then expressed by:

$$S = \frac{D_M - D}{\log_e I} = \frac{1}{G}$$

which yields the optimum value of G. D_M is the value of D for $I = 1$.

The straight line may be termed the impairment characteristic associated with the particular impairment being considered. It will be noted that the straight line can be defined by the characteristic values D_M and G of the logistic function.

3.2 Correction of the scale boundary effect

A kind of "scale boundary effect" has been identified in which observers tend not to use the extreme values of the judgement scale. This may arise from a number of factors, including a psychological reluctance to make extreme judgements, regression to the mean (i.e. centre value) due to the internal variability of the judgement process, and "residual impairment" (even in reference pictures).

In cases in which the measurement involves a difference score (e.g. reference minus test) as in the double-stimulus continuous quality-scale method, this may not present a serious difficulty in interpretation as the effect may be of similar scope in both cases and, thus, cancel.

If it is felt necessary or desirable, the following procedure may be used to adjust scores to cover the full range of the judgement scale.

(NOTE 1 – This correction procedure involves assumptions and can be misleading, so caution is advised in using the procedure; its use should be reported in the presentation of results.)

$$U' = (U_{max} - U_{mid}) \frac{U - U_{mid}}{U_0 - U_{mid}} + U_{mid} \tag{8}$$

where:

U': true score

U: experimental score

U_0: experimental score without distortions

U_{min}, U_{mid}, U_{max}: minimum, mid-value and maximum of range of scores.

For the five-grade category scale and normalized quality scores, the formula is written as follows:

$$u' = \frac{u + u_0 - 1}{2u_0 - 1} \tag{9}$$

3.3 Approximation by a non-symmetrical function

3.3.1 Description of the function

In practice, the use of a symmetrical logistic function frequently induces strong differences between actual data and approximation. These discrepancies may be due to the end of scale effects or simultaneous presence of several impairments in the test which may influence the statistical model and deform the theoretical logistic function. The issue of these complex artefacts is generally a skewness in the function providing the relationship between the mean scores U and the objective measures of the distortion D.

A method to correct some of these artefacts is proposed in § 3.2 but the perfect logistic approximation may rarely be obtained, so, another function is proposed in order to take into account all the parameters. The purpose of this very

simple approximation is reduced to the statistical analysis of the data and not based on an observer's behaviour theory. The function approximates the logistic one in a non-symmetrical way. For a five-grade scale, the formula is:

$$U = \frac{4}{1 + (D_M/D)^{1/G}} + 1$$

the notation being the same as in § 3.1.

If U is normalized as in § 3.1, we obtain:

$$u = \frac{U - 1}{4}$$

and

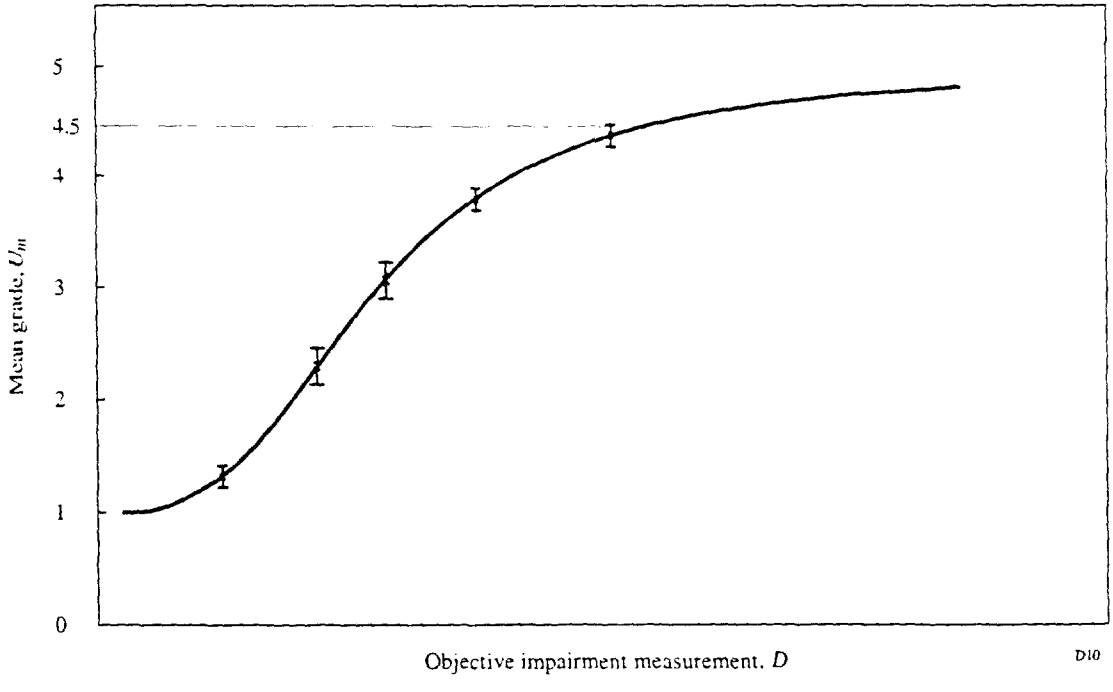
$$u = \frac{1}{1 + (D_M/D)^{1/G}}$$

3.3.2 Estimation of the parameters of the approximation

The estimation of the optimal parameters of the function that provides the minimum residual errors between the actual data and the function may be obtained with any recursive estimation algorithm. Figure 10 shows an example of the use of the non-symmetrical function to represent actual subjective data. This representation allows the estimation of specific objective measures corresponding to interesting subjective value: 4.5 on the five-grade scale, for instance.

FIGURE 10

Non-symmetrical approximation



3.4 Incorporation of the reliability aspect in the graphs

From the mean grades for each impairment tested and the associated 5% confidence intervals, three series of grades are constructed:

- minimum grade series (means – confidence intervals);
- mean grade series;
- maximum grade series (means + confidence intervals).

The estimation parameters for the three series are then estimated independently. The three functions obtained can then be drawn on the same graph, the two from the maximum and minimum series as dotted lines and the mean estimate as a solid line. The experimental values are also plotted on this graph (see Fig. 11). We thus get an estimate of the 5% continuous confidence region.

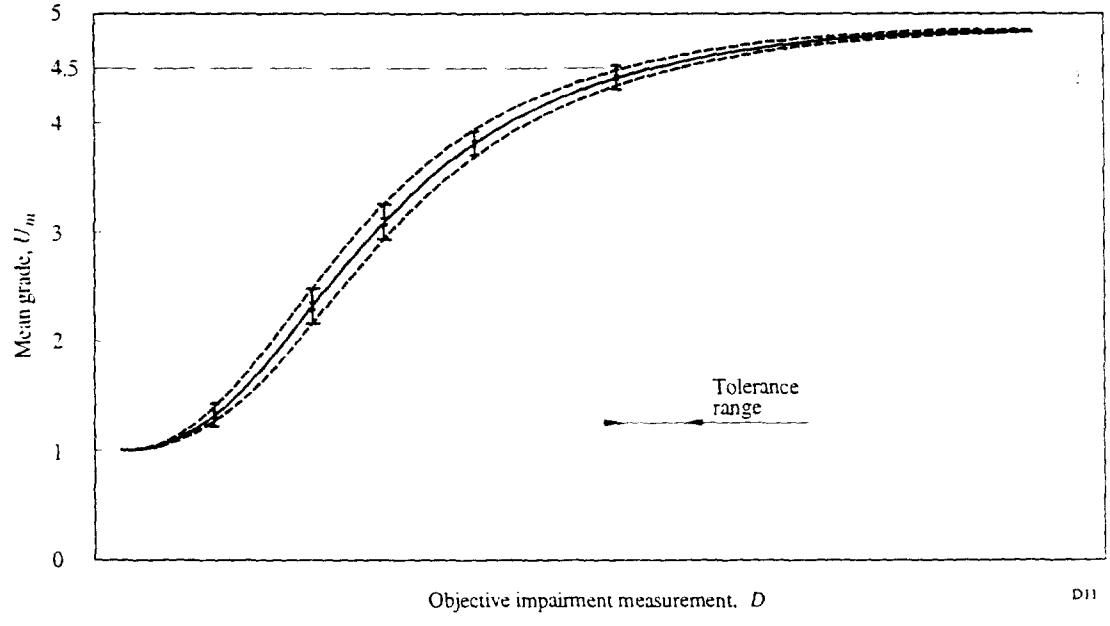
For the grade 4.5 (threshold of visibility for the method) we can thus read off directly from the graph an estimated 5% confidence interval that can be used to determine a tolerance range.

The space between the maximum and minimum curves is not a 5% interval, but a mean estimate thereof.

At least 95% of the experimental values should lie within the confidence region; otherwise it may be concluded that there was a problem in carrying out the test or that the function model chosen was not the optimum one

FIGURE 11

Case of a non-symmetrical impairment characteristic



4 Conclusions

A procedure for the evaluation of the confidence intervals, i.e. the accuracies of a set of subjective assessment tests, has been described.

The procedure also leads to the estimation of mean general quantities that are relevant not only to the particular experiment under consideration, but also to other experiments carried out with the same methodology.

Therefore, such quantities may be used to draw diagrams of the confidence interval behaviour which are helpful for the subjective assessments, as well as for planning future experiments.